Journal of The Korea Society of Computer and Information
Vol. 24 No. 3, pp. 105-112, March 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.03.105

A Table Integration Technique Using Query Similarity Analysis

Go-Bong Choi*, Yong-Tae Woo**

Abstract

In this paper, we propose a technique to analyze similarity between SQL queries and to assist integrating similar tables. First, the table information was extracted from the SQL queries through the query structure analyzer, and the similarity between the tables was measured using the Jacquard index technique. Then, similar table clusters are generated through hierarchical cluster analysis method and the co-occurence probability of the table used in the query is calculated. The possibility of integrating similar tables is classified by using the possibility of co-occurence of similarity table and table, and classifying them into an integrable cluster, a cluster requiring expert review, and a cluster with low integration possibility. This technique analyzes the SQL query in practice and analyse the possibility of table integration independent of the existing business, so that the existing schema can be effectively reconstructed without interruption of work or additional cost.

▸Keyword: SQL similarity analysis, Table integration, Data Architecture, Data Modeling Schema Reconstruction

I. Introduction

기업 데이터베이스는 실무적인 필요에 따라 때때로 새로운 테이블을 생성하거나 칼럼의 추가 등과 같은 테이블의 구조 변경이 발생한다. 그리고 성능 개선을 위하여 반정규화 과정을 수행하기도 한다[1]. 하지만 반정규화는 성능을 향상할 수 있지만 데이터의 무결성을 훼손할 수 있는 새로운 문제점이 발생할 수 있다. 특히 빈번한 테이블 구조의 변경은 데이터 중복을 야기할 수 있으며, 오히려 데이터의 품질을 저하시킬 수 있다[2, 3]. 그리고 반정규화가 반드시 성능 향상에 기여하는 것은 아니라는 연구 결과도 제시되었다[4]. 이러한 중복 데이터로 인한 문제는 데이터 통합을 통하여 효율적인 정보 관리와 데이터 중복으로 인한 오류를 최소화할 수 있다[5].

S. Castano 등은 스키마 클러스터링에 기초하여 참조 개념 스키마를 추출하고, 스키마 유사성을 결정하기 위한 방법을 제시하였다[6]. H. Köpcke 등은 효과적인 데이터 통합을 위한 중요한 작업의 하나인 엔터티 매칭을 위한 다양한 종류의 엔터티 매칭 프레임워크를 비교 분석하였다[7]. 이홍걸 등은 XML을 이용한 데이터베이스 통합 방법과 개체 및 속성 간의 유사도 측정에 기반한 충돌 식별법을 제안하였다[8]. 하지만 기존 연구에서는 스키마 분류를 통하여 유사성을 분석하거나 데이터 통합을 위한 엔터티 매칭 프레임워크를 평가하였지만 실무에서 운용중인 데이터의 통합을 위한 구체적인 방법을 제시하지는 않았다. 또한 XML을 이용한 데이터베이스 통합 방법에 대한 연구[8,9]도 있었지만 실무에서 사용중인 SQL 질의어를 이용한 데이터 통합 방법을 제시한 연구는 진행되지 않았다. 그리고 기존 연구는 기업 내부의 상세한 정보를 파악하기 어려운 상태에서 스키마 변화에 효과적으로 대처하기 어려운 문제점이 있다.

본 연구에서는 업무에서 사용 중인 SQL 질의어 간의 유사도를 분석하여 유사 테이블을 효과적으로 통합하기 위한 기법을 제안하였다. 제안 방법은 전처리 단계, 유사 테이블 군집 생성 및 테이블 동시 출현 분석, 그리고 테이블 통합 지원 단계로 구성된다. 먼저, 전처리 단계에서는 SQL 질의어를 파싱 하여 테이블 이름, 칼럼 이름과 같은 테이블 정보를 추출하는 단계이다. 유사 테이블 군집 생성은 테이블 정보를 이용하여 테이블 간의 유사도를 측정하고, 이를 이용하여 유사 테이블 군집을 생성하는 단계이다. 그리고

• First Author: Go-Bong Choi, Corresponding Author: Yong-Tae Woo
*Go-Bong Choi (gbchoi@coreerp.co.kr), Core Information Technology
**Yong-Tae Woo (ytwoo@changwon.ac.kr), Dept. of Computer Engineering, Changwon National University
• Received: 2019. 02. 08, Revised: 2019. 03. 25, Accepted: 2019. 03. 25.
• This paper has been studied by the project for the exchange professor of Changwon University in 2017.

테이블 동시 출현 분석은 SQL 질의어에서 테이블이 출현한 빈도수와 테이블 간의 동시 출현 빈도수를 측정하여 테이블의 동시 출현 확률을 계산하는 과정이다. 마지막으로 테이블 통합 지원 단계는 유사 테이블 군집과 테이블 동시 출현 확률을 이용하여 테이블의 통합 가능성을 판단하고, 유사 테이블의 통합을 지원하는 단계이다.

본 연구에서 제안한 기법에 대한 효율성을 검증하기 위하여 기업에서 업무용으로 사용하는 15,000여 개의 SQL 질의어를 이용하여 테이블 통합 실험을 진행하였다. 먼저, 테이블 정보로부터 테이블 유사도를 측정하여 유사 테이블 군집을 생성하였다. 그리고 SQL 질의어에서 테이블의 동시 출현 확률을 분석하였다. 유사 테이블 군집에서 테이블 동시 출현 확률을 고려하여 통합 가능성이 높은 군집, 전문가 검토가 필요한 군집 그리고 통합 가능성이 낮은 군집으로 분류하였다. 제안 기법은 컨설팅에 따른 업무 중단이나 기업 내부의 상세한 정보를 파악하지 않고도 유사 테이블 간의 통합을 통하여 스키마 재설계를 효과적으로 지원할 수 있는 기법이다.

II. Preliminaries

1. Related works

1.1 SQL query similarity measurement

SQL 질의어 간의 유사도를 측정하기 위한 기존 연구는 SQL 질의어를 SELECT, FROM, WHERE 절로 구분하여 유사도를 측정하는 방법이 대부분이다. D. P. Groth 는 SQL 질의어의 구조적인 유사도를 관계 대수를 이용하여 측정하고, 그래프를 이용하여 시각화하는 기법을 제안하였다[10]. 한윤희는 유사질의 매칭 기법을 기반으로 데이터베이스의 성능을 향상하기 위한 캐시 엔진을 설계하고 구현하였다[11].

1.2 Silhouette coefficient

실루엣 계수는 각 개체가 군집 내에 얼마나 잘 포함되었는지를 판별하는 방법이다[12]. 실루엣 계수는 대상 객체의 군집에 존재하는 다른 객체와의 평균 거리와 다른 군집 간의 평균 거리를 통해 측정된다. 실루엣 계수의 범위는 -1에서 1까지로 1에 가까울수록 군집화가 잘 되었음을 의미하고, 0은 반대의 의미이다. 그리고 음수 값은 개체가 잘못 분류된 것을 의미한다. 평균 실루엣 계수가 최대 값을 가질 때 적절한 군집화가 되는 성질을 이용하여 군집화 과정에서 군집 수를 잘 모르는 경우에도 최적의 군집 수를 찾는데 적용할 수 있다.

III. Table Integration Model

본 연구에서는 기업에서 업무용으로 사용하는 SQL 질의어 간의 유사도를 분석하여 유사 테이블을 효과적으로 통합할 수 있는 기법을 제안하였다. 그림 1은 본 연구에서 제안한 모델에 대한 개념도이다. 그림 1과 같이 제안 모델은 크게 전처리 단계, 분석 단계 그리고 테이블 통합 지원 단계로 구성된다. 먼저, 전처리 단계는 SQL 질의어에서 테이블 정보를 추출하기 위하여 질의어를 파싱하는 단계이다. 분석 단계는 테이블의 통합 가능성을 판단하기 위한 정보를 생성하는 단계이다. 테이블 통합 지원 단계는 유사 테이블 군집과 테이블 동시 출현 확률을 고려하여 테이블의 통합 가능성을 판단하기 위한 단계이다.

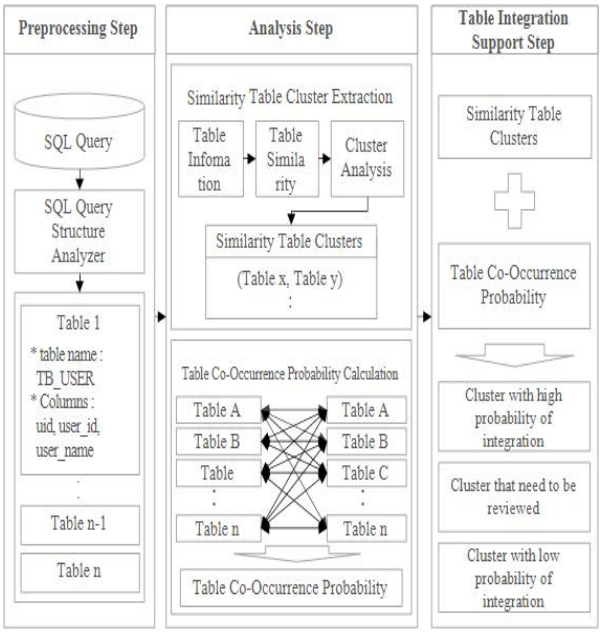

Fig. 1. Conceptual diagram for table integration support model

1. Preprocessing stage

전처리 단계는 질의어로부터 테이블 정보를 추출하기 위하여 SQL 질의어를 파싱하는 단계이다. 먼저, 불완전한 SQL 질의어를 제거하고, SQL 질의어 구조 분석기를 통해 테이블 이름과 칼럼 이름을 추출한다. 추출된 칼럼 이름은 테이블 이름과 매칭하여 테이블 정보를 구성한다. 이 단계에서는 FROM 절의 테이블 이름, SELECT 절의 컬럼 이름과 같은 테이블 정보가 구성된다. WHERE 절은 함수 사용이 빈번하여 분석이 어려운 관계로 제외하였다. 그림 2는 전처리 단계에 대한 개념도이다.

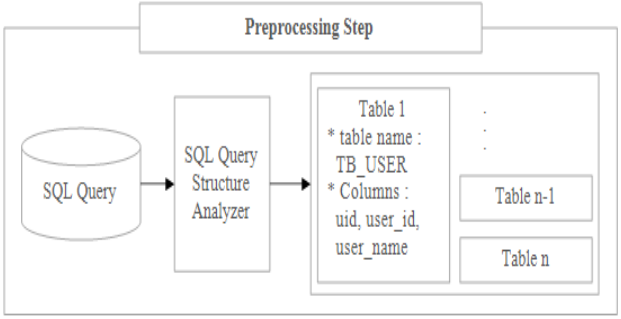

Fig. 2. Conceptual diagram of preprocessing step

2. Anaysis for table integration possibility

분석 단계는 테이블 통합 여부를 판별하기 위하여 필요한 정보를 생성하는 단계이다. 이 단계는 유사 테이블 군집 추출과 테이블 동시 출현 확률 계산 과정으로 구성된다. 먼저, 유사 테이블 군집 추출 과정은 테이블 간의 유사도를 이용하여 군집 분석을 수행하고, 유사 테이블 군집을 생성하는 과정이다. 테이블 동시 출현 확률 계산 과정은 SQL 질의어에서 테이블이 동시에 출현할 확률을 계산하는 과정이다. 유사 테이블 군집과 테이블 동시 출현 확률은 유사 테이블 간의 통합 가능성을 판단하기 위한 정보로 사용된다. 그림 3은 분석 단계에 대한 개념도이다.

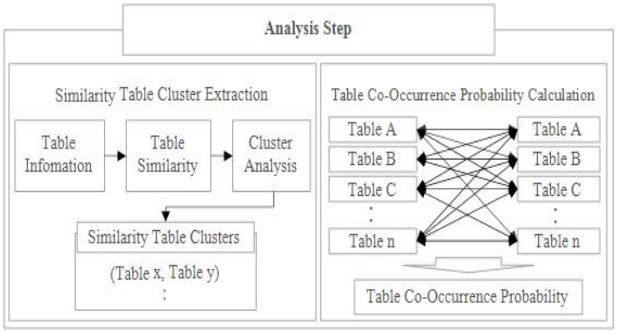

Fig. 3. Conceptual diagram of analysis step

2.1 Extraction of similarity table cluster

유사 테이블 군집 추출은 전처리 단계에서 추출된 테이블 정보를 이용하여 유사 테이블 군집을 생성하는 과정이다. 그림 4는 유사 테이블 군집을 추출하는 과정에 대한 흐름도이다.

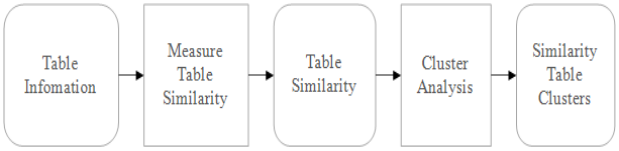

Fig. 4. Flowchart for similarity table clustering

먼저, 전처리 단계에서 추출한 모든 테이블 조합에 대해 테이블 간의 유사도를 측정한다. 그리고 테이블 유사도를 입력 벡터로 하여 계층적 군집 분석 방법에 의해 유사 테이블 군집을 생성한다. 식 (1)은 자카드 지수(Jaccard Index)를 이용하여 두 테이블 간의 유사도를 측정하기 위한 식으로, 두 개체 간의 유사도를 구하는 방법중의 하나이다[13].

$$sim(T_1, T_2) = \frac{n(T_1 \cap T_2)}{n(T_1) + n(T_2) - n(T_1 \cap T_2)} \quad (1)$$

T_1과 T_2는 유사도를 측정할 두 테이블을 의미한다. $T_1 \cap T_2$는 테이블1과 테이블2에 대한 칼럼들의 교집합을 의미한다. $n(x)$는 x 테이블의 칼럼 수를 의미한다. 식 (1)에서 자카드 지수를 이용한 테이블 간의 유사도 측정은 두 테이블 간에 공통되는 컬럼 수를 중복 없는 칼럼 수로 나누어 계산한다. 테이블 유사도는 테이블의 칼럼 수가 3개 이상인 경우에만 측정하였다. 그 이유는 칼럼이 2개 이하인 테이블은 유사도가 높다 하더라도 무의미한 정보일 가능성이 높기 때문이다.

그림 5는 테이블 간의 유사도를 이용한 군집 분석의 수행 과정이다.

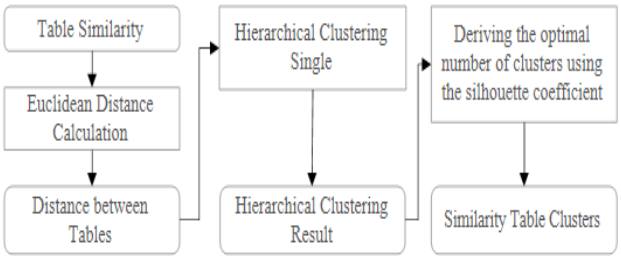

Fig. 5. Procedure for cluster analysis

먼저, 테이블 유사도를 입력 벡터로 하여 테이블 간의 유클리드 거리를 계산한다. 유클리드 거리는 두 점 사이의 최단 거리를 계산할 때 사용하는 방법이다. 유사도가 높은 테이블은 서로 인접하여 위치한 관계로 유클리드 거리가 짧게 나오고, 유사도가 낮은 테이블은 길게 나오게 된다. 그리고 계층적 군집 분석 방법의 최단 연결법(Single Linkage)을 사용하여 테이블들을 군집화하였다. 최단 연결법은 서로 다른 군집에서 가장 가까운 두 점의 거리를 군집 간의 거리로 측정하여 군집을 생성하는 방법이다. 이 방법은 고립된 군집 발견에 유용한 특징이 있다.

마지막으로 계층적 군집 분석 결과에서 실루엣 계수를 측정하여 최적의 군집 수를 찾는다. 최적의 군집 수를 찾기 위하여 군집 수를 2개부터 전체 테이블 개수까지 평균 실루엣 계수를 계산한다. 이 중에서 가장 큰 평균 실루엣 계수를 가지는 경우를 최적의 군집 수로 생성하고, 각 군집에 속한 테이블을 구성한다. 이때, 각 군집은 평균 실루엣 계수를 통해 군집 생성의 적절성 여부를 평가할 수 있다[14]. 표 1은 평균 실루엣 계수를 이용한 군집화의 평가 기준을 의미한다.

Table 1. Evaluation criteria using average silhouette coefficient of cluster

Average Silhouette Coefficient	Clustering evaluation meaning
0.7 ~ 1	Superior clustering results
0.5 ~ 0.7	The center of the cluster is comparative clear
0.25 ~ 0.5	There is noise but can find the center of the cluster
~ 0.25	Virtually impossible to find the center of the cluster

2.2 Calculation for table co-occurrence probability

테이블 동시 출현 확률 계산은 유사 테이블 간의 통합 가능성을 판단하기 위해 테이블 간의 동시 출현 확률을 계산하는

과정으로 다음과 같다. 첫째, SQL 질의어에서 출현한 모든 테이블 정보를 수집한다. 둘째, 테이블 출현 정보에서 테이블이 동시에 출현한 경우에 대한 SQL 질의어 수를 모두 구한다. 셋째, 테이블 출현 빈도수와 테이블 동시 출현 빈도수를 구한다. 넷째, 테이블 동시 출현 확률을 계산한다. 식 (2)는 SQL 질의어에서 두 테이블이 동시에 출현할 확률을 계산하는 식이다.

$$prob(A,B) = \frac{co(A,B)}{n(A)+n(B)-co(A,B)} \qquad (2)$$

식 (2)에서 $n(A)$와 $n(B)$는 A와 B 테이블이 출현한 SQL 질의어의 수를 의미한다. $co(A,B)$는 A와 B 테이블이 동시에 출현한 SQL 질의어의 수이다. 테이블 동시 출현 확률은 유사 테이블 간의 통합 가능성을 판단할 수 있는 기준으로 사용한다. 먼저, 유사 테이블 군집에 속한 테이블들이 SQL 질의어에서 동시에 출현하는 경우가 있는 지를 확인한다. 만일 유사 테이블 군집에 속한 서로 다른 테이블이 통합이 필요할 정도로 유사할 경우, 서로 조인될 필요가 없기때문에 동일한 SQL 질의어에서 테이블이 동시에 출현할 확률이 낮을 것이다. 그리고 유사 테이블 군집의 테이블들이 공통된 테이블에서 동시에 출현하는 경우가 있는 지를 확인한다. 예를 들어, 유사한 테이블 A와 B가 각각의 SQL 질의어에서 공통된 테이블 C와 동시에 출현하여 조인되었다면 테이블 A와 B는 비슷한 용도로 사용됐다고 추정할 수 있다. 따라서 테이블 A와 B는 테이블 통합이 가능한 군집으로 판단할 수 있다.

2.3 Table integration support step

테이블 통합 지원 단계는 유사 테이블 군집과 테이블 동시 출현 확률을 고려하여 테이블 간의 통합 가능성을 판단하기 위한 분석 결과를 제시하는 단계이다. 표 2는 본 연구에서 유사 테이블 간의 통합 가능성을 판단하기 위한 기준이다.

Table 2. Table integration criteria based-on co-occurence of tables in a cluster

Average Silhouette Coefficient	Co-occurrence between tables in a cluster	Tables in a cluster co-occur with the same table	Analysis result
0.7 ~ 1	In all cases	In all cases	Cluster with high probability of integration
0.25 ~ 0.7	X	O	
	O	X	Cluster that need to be reviewed by manually
	O	O	
	X	X	
~ 0.25	In all cases	In all cases	Cluster with low probability of integration

본 연구에서는 평균 실루엣 계수가 0.7 이상인 군집은 통합 가능성이 높은 군집으로 판단하였다. 표 1에 제시된 기준처럼 평균 실루엣 계수가 0.7 이상인 경우에는 군집화가 잘 이루어진 경우이기 때문이다. 그리고 0.25~0.7중에서 군집의 테이블들이 동시에 출현하지 않고 같은 테이블과 동시에 출현하는 경우에는 통합 가능성이 높은 군집으로 판단하였다. 그 이유는 군집내의 테이블들이 같은 테이블과 동시에 출현하는 경우는 해당 테이블들이 비슷한 용도로 사용되었을 가능성이 있기 때문이다. 그리고 군집의 평균 실루엣 계수가 0.25~0.7이면서, 군집의 테이블들이 동시에 출현하거나 군집의 테이블들이 동시에 출현하는 테이블이 없는 군집은 전문가의 검토가 필요한 군집으로 판단하였다. 전문가 검토가 필요한 군집은 통합의 가능성이 있지만 실질적인 통합 여부는 전문가의 실무적인 판단에 따라 통합 여부를 결정할 수 있는 군집이다. 마지막으로 평균 실루엣 계수가 0.25 미만인 군집은 통합 가능성이 낮은 군집으로 통합 고려 대상에서 제외하였다.

IV. Experimental Results

1. Experimental environment

본 연구에서 제안한 질의어 유사도 분석을 이용한 테이블 통합 기법의 효율성을 검증하기 위하여 기업에서 업무용으로 사용중인 15,000여 건의 SQL 질의어를 대상으로 유사 테이블 군집 생성과 테이블 통합 실험을 진행하였다. 군집 분석은 R 프로그램으로 실험하였고, cluster 패키지를 사용하였다.

2. Experimental results

먼저, 전처리 단계에서는 15,000여 개의 업무용 SQL 질의어중에서 604개의 테이블 정보가 추출되었다. 분석 단계에서 유사 테이블 군집은 108개의 군집을 생성했을 때 평균 실루엣 계수가 가장 높았고, 유사 테이블 군집은 22개가 추출되었다. 테이블 동시 출현 확률이 있는 테이블 쌍은 1,504개가 추출되었다. 마지막으로 테이블 통합 단계에서는 통합 가능성이 높은 군집은 9개, 통합 여부에 대한 검토가 필요한 군집은 6개 그리고 통합 가능성이 낮은 군집은 7개로 분석되었다.

2.1 Experimental results for proprocessing step

실험용 SQL 질의어는 기업의 여러 부서에서 오랜기간 동안 실무적으로 사용되어 중복되거나 불완전한 질의어가 다수 존재하는 관계로 전처리 과정을 통하여 해당 질의어를 배제하였다. SQL 질의어 중에서 문법적으로 완전한 질의어는 810개이다. 810개의 질의어에 대하여 SQL 질의어 구조 분석기를 사용하여 테이블 정보를 추출한 결과 총 604개의 테이블이 추출되었다.

2.2 Experimental results for analysis step

분석 단계는 유사 테이블 군집 추출 및 테이블 동시 출현 확률 계산 과정으로 구성된다. 유사 테이블 군집 추출 과정은 자카드 지수를 이용하여 테이블 간의 유사도를 측정하여 유사 테

이블끼리 군집화하는 과정이다. 전처리 단계를 통해 추출한 604개의 테이블로부터 유사도를 측정한 결과 8,200개의 테이블 유사도를 측정할 수 있었다.

표 3은 테이블 간의 유사도가 높은 목록의 일부이다. 604개의 테이블 중에서 테이블 간에 같은 컬럼이 1개라도 있는 152개 테이블을 대상으로 군집 분석을 수행하였다.

Table 3. Part of top list of similar table

Table1	Table2	co-occurrence probability
C1.$L**K	C2.$L**K	1
C1.$L**K	C3.$L**K	1
C2.$L**K	C3.$L**K	1
C4.$L**K	C1.$L**K	1
C4.$L**K	C2.$L**K	1
C4.$L**K	C3.$L**K	1
PN**E	PX***M_D**A	1
P**M_ME**ER	PGR******BER	1
bo**own.V*E_W**K_I**M	V*E_W**K_I**M	0.944
C4.D*V_D*G	C3.D*V_D*G	0.917
PPENV.PP_D*N_P**T_CH***E	PPENV.PP_PR**O_P**T_CH**GE	0.700
ERN.O*T_M*T_VA**E	C2.O*T_M*T_VA**E	0.667
P**M_OB***T	P**M_ACC***OR	0.636
C3.D*V_D*G	C4.D*C_*D	0.615
P**M_ACC***OR	PGR******BER	0.611
PPENV.P*_OB***T	PPENV.P**T_CH**GE	0.600
ADM.org******ion	ADM.Ac**on	0.571
C1.P**T_L**T	C2.P**T_L**T	0.571
C4.D*V_D*G	C4.D*C_*D	0.571
C3.P**T_L**T	C2.P**T_L**T	0.556
C4.E*T_Q*Y	C4.E*T_BA***ON	0.556
ADM.op**on_eff*****ity	ADM.mo**f_eff*****ity	0.500

군집 분석 후, 실루엣 계수를 이용하여 최적의 군집 수를 탐색하였다. 그림 6은 군집 수에 따른 평균 실루엣 계수이다.

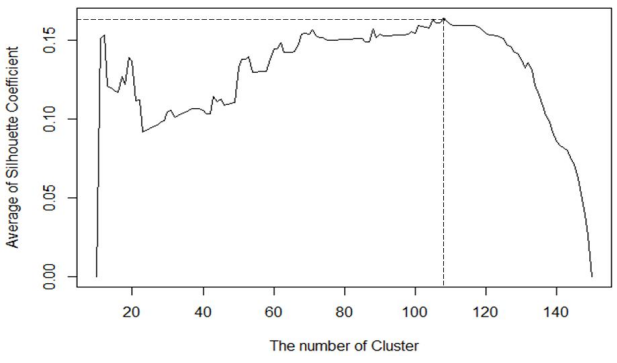

Fig. 6. The number of cluster according to average of Silhouette Coefficient

그림 6에서 X축은 군집 수, Y축은 평균 실루엣 계수를 의미한다. 군집 수를 10개부터 평균 실루엣 계수를 계산하였다. 그 이유는 소수의 군집에서는 다수의 테이블이 집중되는 관계로 군집화의 의미가 없는 상태에서 평균 실루엣 계수가 높게 나타나는 경우가 있기 때문이다. 그림 6에서 군집 수가 108개일 때 평균 실루엣 계수가 최댓값을 가지는 것을 확인할 수 있다. 전체적으로 평균 실루엣 계수가 낮게 나오는 이유는 군집에 속한 테이블이 1개인 군집은 실루엣 계수가 0이 되어 평균값이 낮아지기 때문이다. 108개의 군집 중에서 테이블이 2개 이상인 군집의 평균 실루엣 계수는 0.37이다. 또한 군집 수가 130개 이상일 때 평균 실루엣 계수가 급격히 떨어지는 것을 확인할 수 있다. 그 이유는 군집 수가 늘어날수록 군집에 속한 테이블이 1개인 군집이 많아지는 관계로 평균값이 낮아지기 때문이다.

테이블 동시 출현 확률 계산은 테이블 쌍이 SQL 질의어에서 동시에 사용되는 확률을 계산하기 위한 과정이다. 먼저, 전처리 수행 결과 선별된 810개의 SQL 질의어에서 출현한 모든 테이블 정보를 구하였다. 그리고 SQL 질의어에서 테이블이 출현한 빈도수와 테이블이 동시에 출현한 빈도수를 구하였다. 테이블 출현 빈도수와 동시 출현 빈도수에 의해 테이블 동시 출현 확률을 구한 결과 1,504개의 테이블 쌍이 추출되었다.

표 4는 동시에 출현한 테이블 쌍과 각 테이블이 사용된 SQL 질의어에서 테이블 쌍이 동시에 출현할 확률을 계산한 결과의 일부이다.

Table 4. Part of co-occurrence table pair and co-occurence probability

Table1	Table2	co-occurrence probability
PDA***YPE	PBUS**********FACE	1
ERN.REV***ON	ERN.O*T_M*T_V***E	1
PGRO****MBER	PPOM_ME**ER	0.900
PAPP********ROUP	PMEA**********ROOT	0.571
POPE**********ERTY	PPR***RTY	0.500
PATT****ENTS	PE****SK	0.400
PTYP********TION	PBM*****TION	0.333
P*M_A*L	P*M_A*E	0.250
C1.$L**K	C2.P**T_L**T	0.250
C1.$L**K	C3.P**T_L**T	0.200
C4.$L**K	C1.P**T_L**T	0.200
C4.$L**K	C2.P**T_L**T	0.166
PPOM_G***P	PU**R	0.125

2.3 Table integration support step

테이블 통합 지원 단계는 유사 테이블 군집과 테이블 동시 출현 확률을 이용하여 테이블 간의 통합 가능성을 판단하는 단계이다. 실험에서 유사 테이블 군집은 총 22개로 분석되었다. 통합 가능성이 있는 테이블은 전체 604개의 테이블 중에서 66개로 분석되었다. 본 연구에서 제시한 군집 분석 기준에 의해 통합 가능성이 높은 군집은 총 9개이고, 대상 테이블은 13개이다. 그리고 군집의 평균 실루엣 계수가 0.7 이상인 군집은 2개이고, 0.25에서 0.7 사이의 군집은 7개로 추출되었다. 표 5는 테이블 통합 가능성이 높은 군집의 일부에 대한 예시이다.

Table 5. Sample clusters with high possibility of table integration

Table Name	Cluster Number	Silhouette Coefficient
C3.$L**K	26	1
C1.$L**K	26	1
C4.$L**K	26	1
C2.$L**K	26	1
bo***wn.V*E_W**K_I**M	81	0.938
V*E_W**K_I**M	81	0.940
C4.D*C_*D	25	0.506
C4.D*V_D*G	25	0.730
C3.D*V_D*G	25	0.741
C2.O*T_M*T_VA**E	101	0.651
ERN.O*T_M*T_VA**E	101	0.655
PPENV.PP_D*N_P**T_CH**GE	72	0.596
PPENV.PP_OB**CT	72	0.480
PPENV.P**T_CH**GE	72	0.569
PPENV.PP_PR**O_P**T_CH**GE	72	0.516
ADM.PE**ON_I**O	107	0.528
ADM.PE**ON_K*R	107	0.528

표 6은 통합 가능성이 높은 72번 군집의 테이블 정보이다.

Table 6. Table structure of 72 cluster with a high possibility of table integration

Table Name	PPENV.PP_D*N_P**T_CH**GE	PPENV.PP_PR**O_P**T_CH**GE	PPENV.PP_OB**CT	PPENV.P**T_CH**GE
Column Name	p**t_c*d	p**t_c*d	p**t_c*d	p**t_c*d
	p**t_nu**er	p**t_nu**er	p**t_nu**er	p**t_nu**er
	p**t_ve***on		p**t_ve***on	p**t_ve***on
	int***ace_d**e	int***ace_d**e		int***ace_d**e
	p**t_ac**on	p**t_ac**on		p**t_ac**on
	t**m_*d		t**m_*d	
	t**m		t**m	
	p**t_t**e	p**t_t**e	p**t_n**e	p**t_n**e
	p**t_st**us	p**t_st**us	p**t_e*v	p**t_e*v
	e*n_l**t_s*q	e*n_l**t_s*q	e*n_l**t_s*q	e*n_l**t_s*q

```
SELECT a.p**t_c*d, a.p**t_nu**er, a.p**t_ac**on,
a.e*n_l**t_s*q, a.p**t_t**e, a.p**t_st**us interface_status,
b.c_ma****ty current_status, c.c_3*_n**e, c.c_alt*****ion,
c.d_p**h, c.c_re****on, c.c_res*****ble, c.c_l**t_rep****ory,
c.c_in***rm, c.c_o***r
FROM PPENV.PP_D*N_P**T_CH**GE a, C4.P**T_L**T b,
C4.D*C_*D c
WHERE a.p**t_e*v = 'C4'
AND a.p**t_c*d = b.$**d
AND b.c_ma****ty  = 'RELEASED'
AND b.c_la***od = to_date(:v_d**e,'yyyy-mm-dd')
AND b.$c*d = c.$**d
AND c. C_3*_N**E = :v_file_name;

SELECT a.p**t_c*d, a.p**t_nu**er, a.p**t_ac**on,
a.e*n_l**t_s*q, a.p**t_t**e, a.p**t_st**us interface_status,
b.c_ma****ty current_status, b.p***o, c.$*id, c.$c**d,
c.c_3*_n**e, c.c_alt*****ion, c.d_p**h, c_ve***on
FROM PPENV.PP_PR**O_P**T_CH**GE a, C4.P**T_L**T b,
C4.D*C_*D c
WHERE a.p**t_e*v = 'C4'
AND a.p**t_c*d = b.$**d
AND b.c_ma****ty  = 'RELEASED'
AND b.c_la***od = to_date(:v_d**e,'yyyy-mm-dd')
AND b.$**d = c.$**d
AND c. C_3*_N**E = :v_file_name;
```

Fig. 7. Part of the SQL query where the table in the 72 cluster is used

그림 7은 72번 군집의 테이블이 사용된 2종류의 SQL 질의어이다.

표 6의 72번 군집은 평균 실루엣 계수가 0.54로 표 1의 평가 기준에 의해 비교적 군집의 중심이 명확한 군집이다. 표 6의 테이블 정보에서 모든 컬럼이 다른 테이블의 컬럼과 적어도 1개는 매칭되는 것을 확인할 수 있다. 그리고 그림 7에서처럼 군집 내의 테이블들이 동시 출현하는 경우가 없고, 군집의 테이블들이 공통된 테이블(C4.P**T_L**T b, C4.D*C_D c)과 동시에 출현하는 것을 확인할 수 있다. 따라서 72번 군집은 표 2의 평가 기준에 따라 통합 가능성이 높은 군집으로 분석할 수 있다.

전체 군집 중에서 테이블 통합 여부에 대한 전문가의 검토가 필요한 군집은 총 6개이고, 대상 테이블은 16개이다. 전문가의 검토가 필요한 군집은 평균 실루엣 계수가 0.25에서 0.7 사이인 군집이다. 또한, 검토가 필요한 군집으로 분류하는 동시 출현 조건은 군집 내의 테이블이 서로 동시에 출현하거나 군집의 테이블간에 동시 출현하지 않으면서 군집의 테이블들이 같은 테이블과 동시에 출현하지 않는 경우이다.

표 7은 전문가의 검토가 필요한 군집의 일부 예시이다.

Table 7. Sample clusters that need to be reviewed by expert

Table Name	Cluster Number	Silhouette Coefficient
PGRO*****BER	40	0.642
PPOM_AC****OR	40	0.545
PPOM_ME**ER	40	0.642
PPOM_OB**CT	40	0.251
ADM.Ac**on	103	0.450
ADM.org******ion	103	0.446
C1.P**T_L**T	91	0.427
C2.P**T_L**T	91	0.522
C3.P**T_L**T	91	0.507
C4.P**T_L**T	91	0.442
PMEPR**********ASTER	104	0.417
PME**********TER	104	0.408
PI****RM	53	0.288
PIM*******ION	53	0.285
PFN**********DIT	57	0.246
PFN**********DIT	57	0.281

표 8는 전문가의 검토가 필요한 91번 군집의 테이블 정보이다.

Table 8. Table of 91 cluster that need to be reviewed by manually

Table Name	C1.P**T_L**T	C2.P**T_L**T	C3.P**T_L**T	C4.P**T_L**T
Column Name		OP**ON_C**E	e*_c**e	p**t_*o
		p**t_ve***on	p**t_ve***on	p**t_ve***on
				p**t_n**e
	ve***on	ve***on	ve***on	ve***on
	co**t	co**t	co**t	co**t
	cr**te_u**r	cr**te_u**r	cr**te_u**r	cr**te_u**r
	la***od_u**r	la***od_u**r	la***od_u**r	la***od_u**r
	MA****TY		c_ma****ty	c_ma****ty
			e*_n**e	pr**o

그림 8은 91번 군집의 테이블이 사용된 SQL 질의어에 대한 예이다.

```
  SELECT b.OP**ON_C**E "parent", b.p**t_ve***on
"p_ver",c.e*_c**e "child", c.p**t_ve***on "c_ver",
a.$**d child_cid, a.$c**d child_cpid, 'C3' child_env,
a.$co**me, a.$cp**me, a.$t**e, c.ve***on, c.co**t,
c.cr**te_u**r, c.la***od_u**r
  FROM C3."$L**K" a, C2.P**T_L**T b, C3.P**T_L**T c
  WHERE a.$**d = b.$**d
  AND a."$**d_R*F" = c.$**d
  AND a.$**d = hextoraw(:v1)
  AND a.$c**d = hextoraw(:v2);
```

Fig. 8. Part of the SQL query where the table in the 91 cluster is used

91번 군집의 평균 실루엣 계수는 0.47로 표 1의 기준에 의해 노이즈가 있지만 군집의 중심을 찾을 수 있는 군집이다. 즉, 이 군집은 통합 가능성이 높은 군집의 테이블보다 매칭되는 컬럼이 적다. 그리고 그림 8에서 군집에 속한 테이블(C2.P**T_L**T b, C3.P**T_L**T c)들이 동시에 출현하는 것을 확인할 수 있다.

전체 군집 중에서 통합 가능성이 낮은 군집은 7개로 분석되었고, 대상 테이블은 27개이다. 이 군집은 평균 실루엣 계수가 0.25 보다 낮은 관계로 군집의 중심을 찾는 것이 거의 불가능한 군집들이다. 이 군집의 평균 실루엣 계수가 낮게 나오는 이유는 자기 군집의 개체 간의 평균 거리가 다른 군집의 개체와의 평균 거리와 비슷하기 때문이다.

V. Conclusions

본 논문에서는 SQL 질의어의 유사도를 분석하여 테이블 간의 통합 여부를 지원하기 위한 기법을 제안하였다. 먼저, SQL 질의어 구조 분석기를 통해 SQL 질의어를 파싱하여 테이블 정보를 추출하였다. 추출된 테이블 정보로부터 테이블 유사도를 측정하였다. 이를 이용하여 계층적 군집 분석 방법에 의해 유사 테이블 군집을 생성하였다. 그리고 SQL 질의어에서 테이블들이 동시에 출현하는 확률을 계산하였다. 그리고 유사 테이블 군집과 테이블 동시 출현 확률을 분석하여 통합이 필요한 군집, 전문가의 검토가 필요한 군집, 통합 가능성이 낮은 군집으로 분류하여 유사 테이블 간의 통합 가능성을 분석하였다.

제안 기법의 효율성을 검증하기 위해 기업에서 수 년동안 업무용으로 사용하는 SQL 질의어를 대상으로 실험하였다. 전체 SQL 질의어중에서 불완전한 SQL 질의어를 제거하고, 810개의 SQL 질의어를 대상으로 실험하였다. 분석 단계에서는 파싱된 604개의 테이블을 대상으로 테이블 유사도를 측정하였다. 계층적 군집 분석을 실시한 결과 108개의 군집이 생성되었고, 유사 테이블 군집은 22개가 추출되었다. 테이블 통합 지원 단계에서 군집 분석 기준을 통해 통합 가능성이 높은 군집은 9개, 전문가 검토가 필요한 군집은 6개 그리고 통합 가능성이 낮은 군집은 7개로 분석되었다.

제안 기법은 현업에서 업무용으로 사용 중인 SQL 질의어를 이용하여 진행 중인 업무와 독립적으로 테이블 통합 가능성을 효과적으로 판단할 수 있다. 따라서 컨설팅에 따른 업무 중단이나 별도의 비용 지출 없이 유사 테이블의 통합을 지원하여 기존 스키마의 재구성이 가능하다. 그리고 유사 테이블로 인해 발생할 수 있는 데이터 중복을 개선하여 데이터의 무결성을 향상시킬 수 있다.

REFERENCES

[1] Jong Suk Lee, Chang Ho Lee, “Modeling on Data Performance for Very Large Database,” Proceedings of the Korea Safety Management & Science, No. 1, pp. 383-391, 2012.

[2] Kang Soo Seo, “*The Guide for Data Architecture Professional*” Korea Data Agency, pp. 214-553, 2013.

[3] R. Y. Wang, V. C. Storey and C. P. Firth, “A Framework for Analysis of Data Quality Research,” Transactions on Knowledge and Data Engineering, Vol. 7, No. 4, pp. 623-640, Aug. 1995.

[4] Hae Kyung Rhee, “Harmfulness of Denormalization Adopted for Database for Database Performance Enhancement,” Journal of the Institute of Electronics and Information Engineers, Vol. 42, No. 3, May 2005.

[5] Hye Young Seo, Seo Young Kwon, Jae Kwon Ahn, Young Jin Kim, “A Case Study on the Implementation of Master Data Management System for Global Manufacturing Company,” Entrue Journal of Information Technology, Vol. 7, No. 2, pp. 91-102, Jul. 2008.

[6] S. Castano, V. Antonellis, M. G. Fugini and B. Pernici, “Conceptual Schema Analysis: Techniques and Applications,” ACM Transactions on Database Systems, Vol. 23, No. 3, pp. 286-333, Sep. 1998.

[7] H. Köpcke and E. Rahm, “Frameworks for entity matching: A comparison,” Data & Knowledge Engineering, Vol. 69, No. 2, pp. 197-210, Feb. 2010.

[8] Hong Girl Lee et al., “A Study on the Database Integration Methodology using XML,” Journal of Korean Navigation and Port Research, Vol. 29, No. 5, pp. 883-890, Dec. 2005.

[9] Sanjay Madria et al., “An XML Schema Integration and Query Mechanism System,” Data & Knowledge Engineering, Vol. 65, No. 2, pp. 265-303, May 2008.

[10] D. P. Groth, “Visual Representation of Database Queries using Structual similarity,” Information Visualization, pp.

102-107, 2003.
[11] Yun Hee Han, "*Design and Implementation of Database Cache engine based on Similarity Query Matching*" Masterr's Thesis, Korea Polytechnic University, 2008.
[12] P. J. Rousseeuw, "Silhouette: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, Vol. 20, pp. 53-65, Nov. 1987.
[13] Soojung Lee, "Performance Analysis of Similarity Reflecting Jaccard Index for Solving Data Sparsity in Collaborative Filtering," The Journal of Korean Association of Computer Education, Vol. 19, No. 4, pp. 59-66, Jul. 2016.
[14] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, New York, 1990.

Authors

Go-Bong Choi received the B.S. and M.S. degrees in Computer Science and Engineering from Changwon National University, Korea, in 2015 and 2017 respectively. Mr. Choi is a Manager in the Technical support team Core Information Technology since 2017. He is interested in Data Modeling, ERP System.

Yong-Tae Woo received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Kyungpook National University, Korea, in 1982, 1984 and 1995, respectively. Dr. Woo is a Professor in the Department of Computer Engineering, Changwon National University since 1987. He is also CEO of Hibrain.net Co. He is interested in Data Modeling, Internet Business, and Big Data Analysis areas.